AMD RETAIL AI SOLUTIONS

THE TIPPING POINT FOR RETAIL AI IS HERE

Shrinking AI overhead and rising CPU performance are making in-store retail AI viable at scale.

AI Inference costs have **dramatically dropped** — **280X**[1]

Model sizes have **decreased significantly** — **142X**[2]

Lightweight AI is doing **heavy duty work**

Falling costs and shrinking models are making AI viable on standard, off-the-shelf servers.

AI IS ALREADY TRANSFORMING RETAIL

63%
of senior retail managers think AI is key to staying competitive.[3]

70%
of retailers use AI in marketing.[4]

54%
of retailers use AI in merchandising, strategy, and pricing.[4]

58%
of consumers have used smart carts, digital kiosks, or enhanced self-checkouts.[3]

MOVE FAST WITH AMD RETAIL AI SOLUTIONS

Get a jumpstart on your AI efforts with tested and validated AI solutions from our open ecosystem of retail AI developers and manufacturers.

REINVENT IN-STORE OPERATIONS WITH INTELLIGENT EDGE-CLASS SERVERS

Servers featuring AMD EPYC™ 8534PN Server CPUs deliver 64 cores of AI-capable performance in a 175W to 200W power envelope — an ideal profile for the power-constrained realities of in-store servers, video systems, and point of sale hardware.

AMD EPYC™ 8534PN SERVER CPUs DELIVER THE PERFORMANCE AND EFFICIENCY RETAIL AI NEEDS

~58%
higher total AI/ML throughput[5]

~52%
more performance per watt, per dollar[5]

~42%
lower system power[5]

AMD EPYC 8534PN Server CPUs deliver higher AI throughput than Xeon Platinum 8471N at retail-friendly cost and power budgets. See note 5 for details.

Deploy at a global scale with industry leaders

Lenovo   HPE   SUPERMICRO

7 OF THE 10 LARGEST FORTUNE 100 RETAILERS RUN ON AMD EPYC™ SERVER CPUs[6]

IT'S TIME TO JOIN THEM

Transform your operations with affordable AI running on the same high-performance AMD EPYC Server CPUs trusted by the world's largest retailers. Contact AMD Retail AI Solutions to get started.

[Learn more]

1. The cost of performing inference with a GPT-3.5-sized, ~175 billion-parameter model dropped 280 times since 2022. For details, see page four in the Artificial Intelligence Index Report 2025, from Stanford University's Human-Centered Artificial Intelligence Institute.
2. By 2024, Microsoft's Phi-3-mini matched the 2022 performance of Google's 540 billion-parameter PaLM model using only 3.8 billion parameters, a 142X drop in two years. For details, see page 95 in the Artificial Intelligence Index Report 2025, from Stanford University's Human-Centered Artificial Intelligence Institute.
3. Everseen Insights, "The Future of AI in Retail." February 2025.
4. BRG, "AI in Retail: In Pursuit of Meaningful AI Adoption." November 12, 2025.
5. SP5-006: AI/ML workloads (Neural Magic DeepSparse, oneDNN, OpenCV, OpenVINO, and TensorFlow averaged) performance/system W/system $ comparison based on Phoronix Test Suite paid testing as of 8/18/2023. Configurations: 1P 64C EPYC 8534PN (0.96x relative performance, 247 avg system W, est $8,482 system cost USD) powered server versus 1P 52C Xeon Platinum 8471N (1.07x relative performance, 423 avg system W, est $8,477 system cost USD) powered server for 0.89x the performance, 42% lower system power (1.52x the performance/system W), comparable system cost for 1.52x the overall system performance/W/$. Assuming an 8kW rack deploying servers, 32 ea. EPYC 8534PN vs. 18 ea. Xeon 8471N can fit within the power budget delivering 1.58x the total AI/ML throughput/rack on average. Testing not independently verified by AMD. Scores will vary based on system configuration and determinism mode used (default TDP power determinism mode profile used). This scenario contains many assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. Results may vary and are normalized to EPYC 8534P Performance Determinism mode (always 1 in all measurements). Estimated system pricing based on Bare Metal Server GHL TCO v5.52. https://www.amd.com/en/legal/claims/epyc.html#q-SP5-006.
6. EPYC-059: Top 10 U.S. retail companies by revenue according to 2025 Fortune 500 list as of June 2, 2025. https://fortune.com/ranking/fortune500/. https://www.50pros.com/fortune500. "Fortune 100" refers to the top 20% ranked companies in the 2025 Fortune 500 list, published in June 2025. From Fortune Magazine. ©2025 Fortune Media IP Limited. All rights reserved. Used under license. Fortune and Fortune Media IP Limited are not affiliated with, and do not endorse products or services of Advanced Micro Devices, Inc.

© 2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and other countries. Other product names used in this publication are for identification purposes only and may be trademarks of their respective owners.